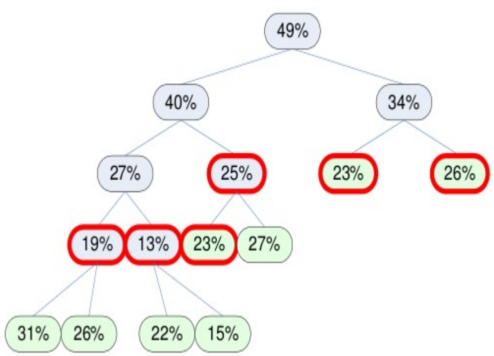
Subspace Indexing on Grassmannian Manifold for Large Scale Visual Analytics

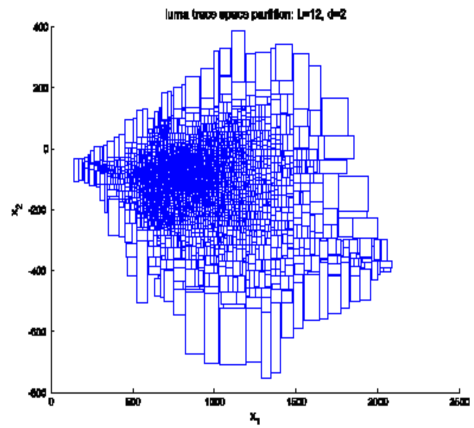
Zhu Li

Media Networking Lab

FutureWei (Huawei) Technology, USA

Bridgewater, NJ





Outline

- **Short Self-Intro**
- **Large Scale Visual Analytics**
  - Applications
  - Key Technical Challenges
  - Query-Driven Local Subspaces
  - Indexed Subspaces on Grassmannian Manifold
  - Simulation
  - Conclusion & Future Work

About Me: http://users.eecs.northwestern.edu/~zli

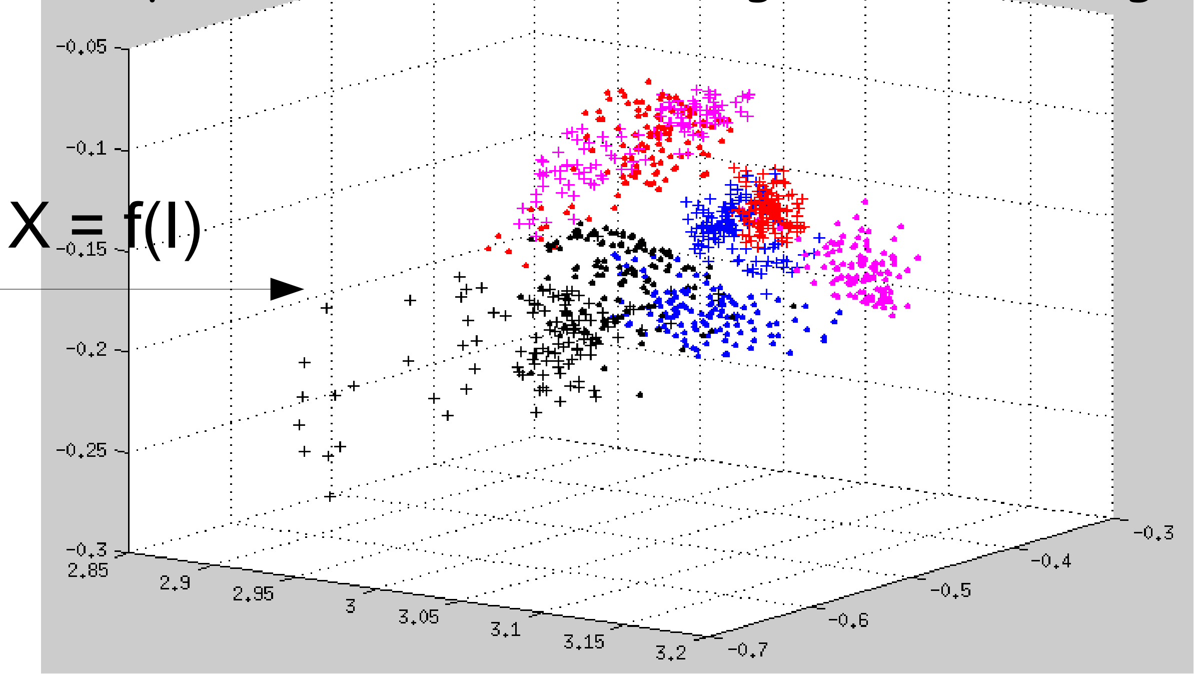
- **Bio:**
  - Media Analytics Group Lead, Core Networks R&D, ***Huawei Tech*** USA, 2010.10~ to date
  - Asst Prof, ***HK Polytechnic Univ***, 2008.04~2010.09
  - Senior, Senior Staff, and then Principal Staff Researcher, Multimedia Research Lab, ***Motorola Labs***, USA, 2000-08.
  - Software Engineer, CDMA Network Software Group***, Motorola CIG***, USA, 1998-2000.
  - PhD in Electrical & Computer Engineering, ***Northwestern University***, USA, 2004.
- **Research Interests:**
  - Large scale audio/visual data analysis, storage and indexing, search and mining.
  - Video Adaptation, Image/Video QoE Modelling, Very Low Bit Rate Video
  - Optimization and distributed computing for Content Delivery Networks (CDN).

The Large Scale Visual Analytics Problems

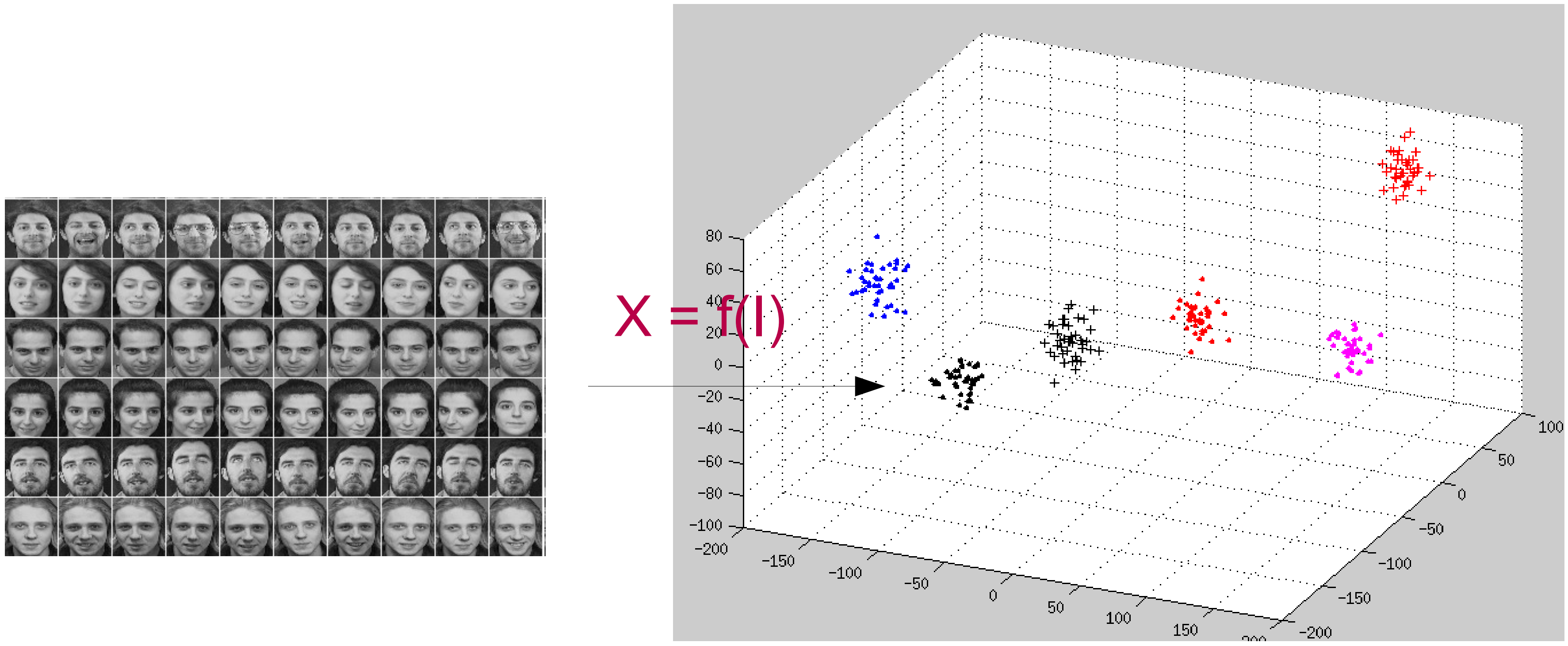
- **Face Recognition**
  - Identify face from 7 million HK ID face data set
- **Image Search**
  - Find out the category of given images

The Problem

- **Identification**
  - Given a set of training image data and label$\{f_k, l_k\}$, and a probe *p*, identify the unique label associated with *p*.
- **Why is it difficult ?**
  - When the number of unique labels, *m*, and training data *n* are large....



- **Find a "good" f()**
  - Such that after projecting the appearance onto the subspace, the data points belong to different classes are easily separable



Global Linear LPP Models: f(X) = AX

- **LPP (Xiaofei He, et.al):**
  - Minimizing weighted distance (a graph) after projection

$$\min_{A} \sum_{j,k} w_{j,k} ||Ax_j - Ax_k||^2$$

-Solve by:

$$XLX^T A = \lambda XDX^T A, s.t. L = D - W, D_{k,k} = \sum_j w_{j,k}$$

- Embed a graph with pruned edges

$$\begin{cases} w_{j,k} = e^{-\alpha||x_j - x_k||}, & \text{if } ||x_j - x_k|| \leq \epsilon \\ 0, & else \end{cases}$$

Global Linear LDA Models: f(X)=AX

- **LDA:**
  - Maximizing inter-class scatter over intra

$$A = \arg\max_{A} |A^T S_B A|, \; s.t. \; |A^T S_W A| = 1$$

$$s_B = \sum_{k=1}^{n} n_k (\overline{X}_k - \overline{X})(\overline{X}_k - \overline{X})^T \qquad s_W = \sum_{k=1}^{n} \sum_{P(X_j)=k} (X_j - \overline{X}_k)(X_j - \overline{X}_k)^T$$

-Solve by:

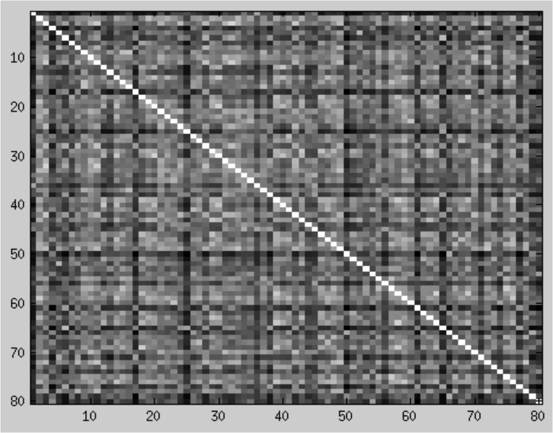
$$S_B A = \lambda S_W A$$

- Embedding a graph with no edges among inter-class points

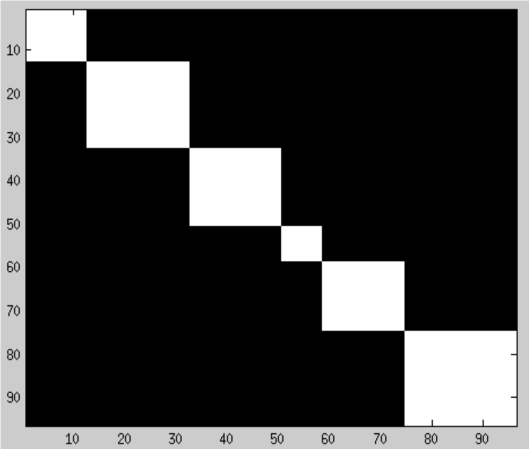
$$\begin{cases} w_{j,k} = \frac{1}{m_i}, & \text{if } x_j, x_k \in \text{class i} \\ 0, & else \end{cases}$$

Graph Embedding Interpretation

- **Find the best embedding**
  - LDA:
    - » preserve the affinity matrix that has zero affinity for data points pairs that are not belonging to the same class
  - LPP:
    - » Have more flexibility in modeling affinity w_{jk}.



LPP Affinity



LDA Affinity

Non-Linear Models

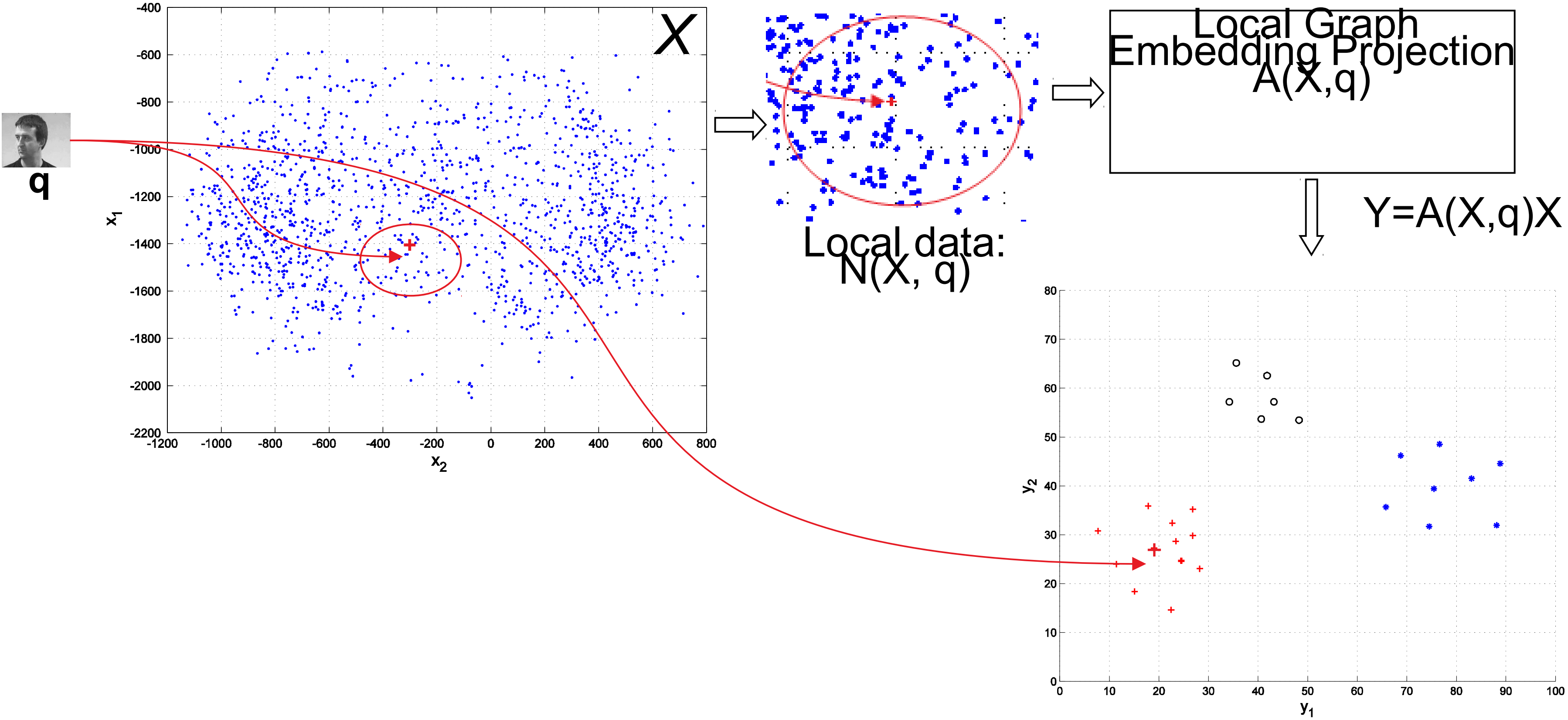
- **Appearance manifolds are non-linear in nature**
  - Global linear models will suffer

- **Non-Linear Solutions:**
  - Kernel method: e.g K-PCA, K-LDA, K-LPP, SVM
    - » Evaluate inner product $<x_j, x_k>$ with a kernel function $k(x_j, x_k)$, which if satisfy the conditions in Mercer's Theorem, implicitly maps data via a non-linear function.
    - » Typically involves a QP problem with a Hessian of size $n \times n$, when n is large, not solvable.
  - LLE /Graph Laplacian:
    - » An algorithm that maps input data $\{x_k\}$ to $\{y_k\}$ that tries to preserve an embedded graph structure among data points.
    - » The mapping is data dependent and has difficulty handling new data outside the training set, e.g., a new query point

- **How to compromise ?**
  - Piece-wise Linear Approximation

Piece-wise Linear : Query Driven

- **Query-Driven Piece-wise Linear Model**
  - No pre-determined structure on the training data
  - Local neighborhood data patch identified from query point *q*,
  - Local model built with local data, *A(X, q)*



Local Model Discriminating Power Criteria

- **What is a good N(X, q) ?**

- **Model power:**
  - A: Dxd, D=wxh

- **Data Complexity: Graph Embedding Interpretation:**
  - PCA: a fully connected graph
  - LDA: a graph with ε-nn edges pruned for intra-class points
  - LPP/LEA; k-nn/ pruned graph
  - as number of edges/relationship among data points

$$|E(X)| = \begin{cases} \binom{n}{2}, & PCA \\ \sum_{j=1}^{m} \binom{n_j}{2}, \ s.t. \sum_{j=1}^{m} n_j = n, & LDA \\ nK, & LPP/LEA \end{cases}$$

- **What is a good compromise of data complexity and model power ?**

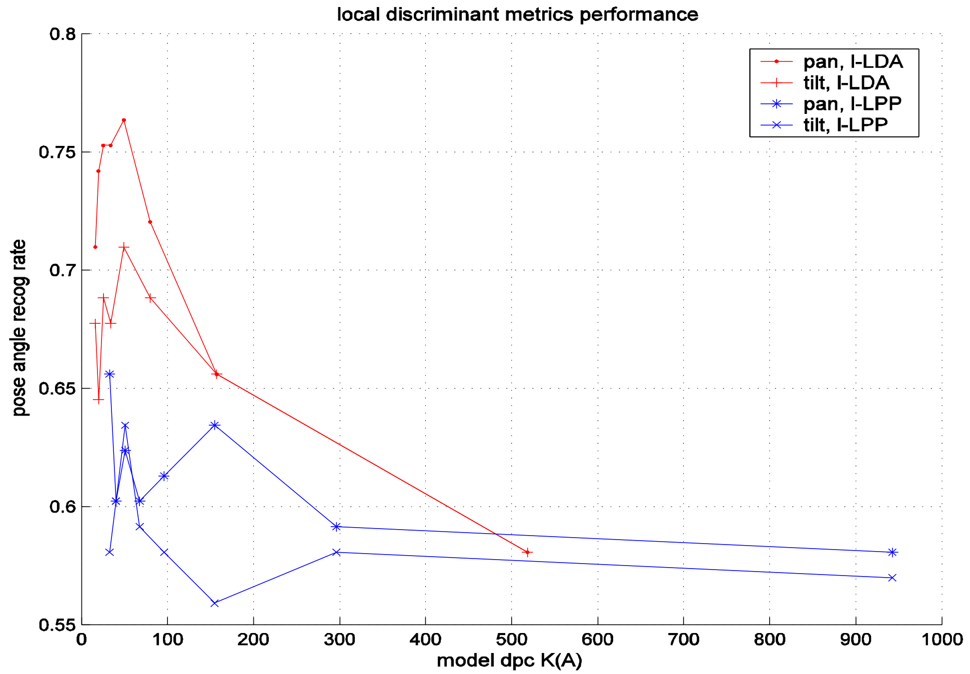
Discriminant Power Co-efficient (DPC)

- **Given the model power constraint:**
  - *w*, *h*, appearance model luminance field size
  - *d*, dimensionality of *A*(x, q)
- **How to identify a neighborhood to achieve a good balance of data complexity and model power ?**

  - DPC, K(A(X,q)) =

$$\frac{w \times h \times d}{|E(X_{(q)})|}$$

  - Need to balance DPC with info loss in node/edge pruning



Head Pose Recognition Performance

- **Recognition rate is improved:**
  - W=18, h=18, K=30

Table 1. Pose estimation error rates

	Pan (d=16)	Tilt (d=16)	Pan (d=32)	Tilt (d=32)
PCA	*33.5*	*44.3*	*26.9*	*35.1*
LDA	*30.1*	*33.3*	*25.8*	*26.9*
LPP[(1)]	*30.1*	*31.2*	*24.7*	***22.6***
LPP[(2)]	*67.7*	*76.3*	*63.4*	*61.3*
l-PCA	*25.2*	*37.8*	*24.5*	*37.6*
l-LPP	*33.9*	*44.5*	*29.2*	*40.2*
l-LDA	***20.4***	***30.7***	***19.1***	*30.7*

- **And the cost in computation is rather modest**
  - Matlab code, online local model A(X,q) learning and NN classification:

Table 2. Computational complexity (sec) per recognition

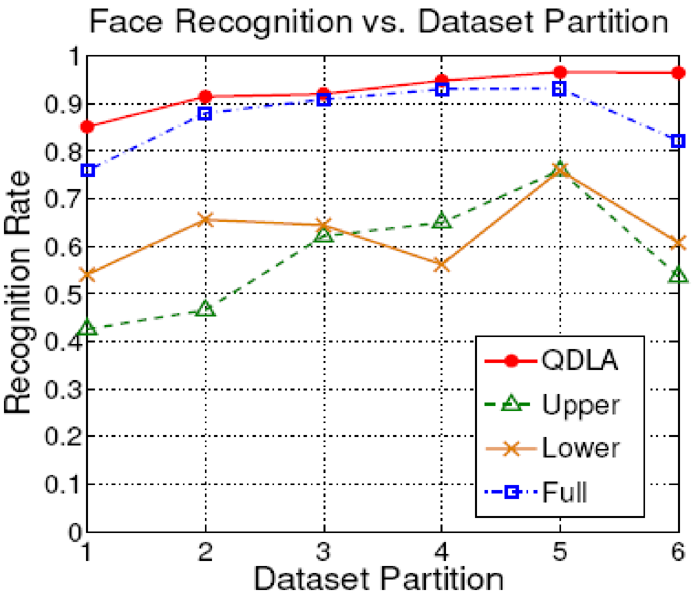
	K=30	*K=60*	*K=90*
l-LDA, $d=16$	*0.105*	*0.132*	*0.121*
l-LDA, $d=32$	*0.145*	*0.146*	*0.176*
l-LPP, $d=16$	*0.094*	*0.122*	*0.104*
l-LPP, $d=32$	*0.132*	*0.116*	*0.144*

Face Recognition Performance

- **Local model combination in face recognition**
  - Query point drives 3 local models, $A_1(X, q)$, $A_2(X, q)$, $A_3(X, q)$
  - Local model classification error estimation,
  - Combining the results – weighted voting

(a) Upper Face Model (b) Lower Face Model (c) Full Face Model

Multiple face models with different area and scale:
(a) Upper face model (18 × 16).
(b) Lower face model (14 × 18).
(c) Full face model (21 × 28).



ORL data set test: leave 1,2,3 out:

Query Driven Solution Problems

- **Optimality of the Local Model is not established**
  - Parameters $\epsilon - NN$, k-NN, and heat kernel size determines the number of non-zero affinity edges in local graph
  - The choice is based on DPC, which is still heuristic
- **Computational Complexity**
  - Need to compute a nearest neighbor set and its affinity, as well as the local embedding model at run time.
  - Need extra storage to store all training data, because the local NN data patch is generated at run time, as function of the query point.
  - Indexing/Hashing scheme to support efficient access of training data.

Stiefel and Grassmannian Manifolds

- **Stiefel manifolds**
  - All possible p-dimensional subspaces in d-dimensional space, A_{pxd}, spans Stiefel Manifold, S(p, d) in R^{dxp}, d > p.

$$\mathcal{S}(p,d) = \{A \in R^{d \times p}, s.t. A'A = I_d\}$$

  - The DoF is not pxd, rather: pd – (1/2)d(d+1)

- **Grassmannian manifolds**
  - G(p, d) identifies p-dimensional subspaces in d-dimensional space
  - It is stiefel manifolds but with an equivalence constraint:
    - » A1 = A2, if span(A1) = span(A2), or
    - » Exist othonormal dxd matrix R_d, A1=A2R_d.
  - The DoF: pd-d^2. G(p, d) is the quotient space of S(p, d)/O(d)

Subspaces on Grassmannian Manifold

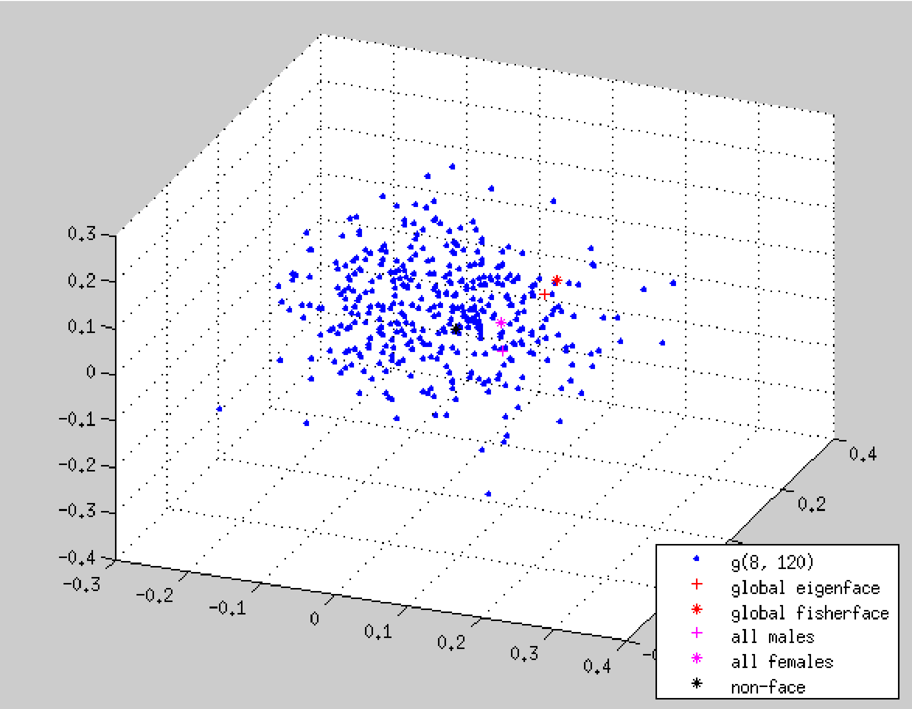
- **The BEST subspace for identification ?**
  - All possible p-dimensional subspaces in d-dimensional space, A_{pxd}, spans Grassmannian Manifold, G(p, d) in R^{dxp}, d > p.
    - eg., G(2, 3), biz card example
  - The DoF of A is not pxd, as for,

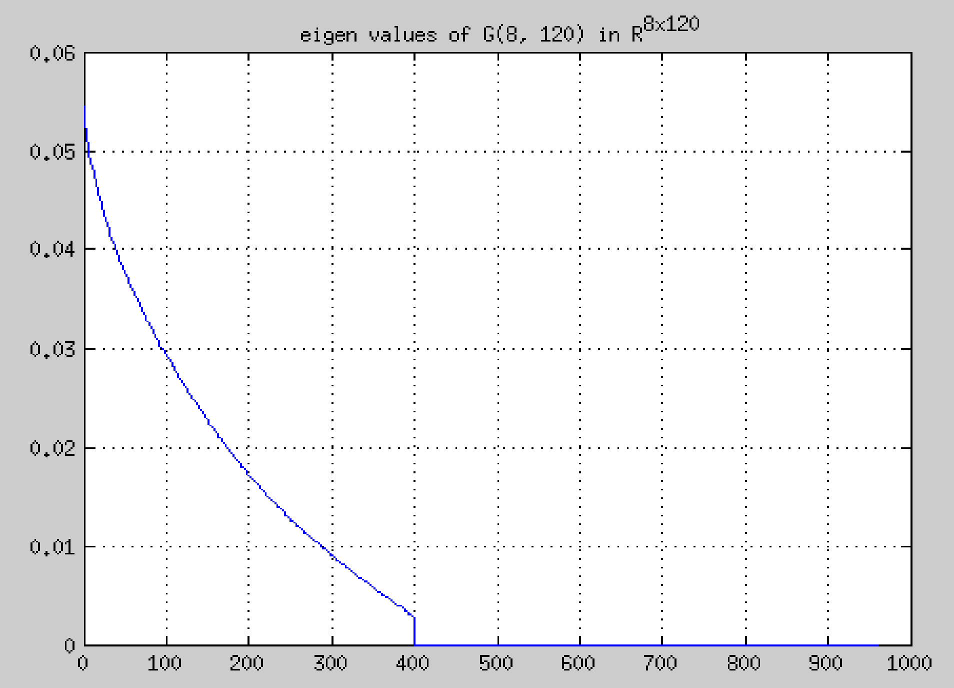
$$< a_j, a_k >= 0, < a_j, a_j >= 1, \text{for } A^T = [a_1, a_2, ..., a_p],$$

  - Face Appearance model, typically, d=400~500, p=10~30.
  - The BEST subspace A* is somewhere on G(p, d), therefore it is important to figure out a way to characterize the similarity between subspaces in G(p, d), and give a structure of all subspace w.r.t the task of identification.

Grassmannian Manifold Visualization

- **Consider a typical appearance modeling**
  - Image size 12x10 pel, appearance space dimension d=120, model dimension p=8.
  - 3D visualization of all S(8, 120) and their covariance eigenvalues"
  - Grassmann Manifolds are quotient space S(8, 120)/O(8)



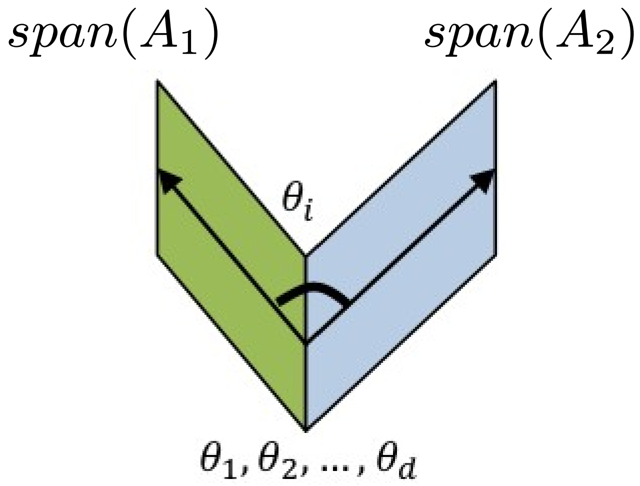


Principle Angles

- **The principle angles between two subspaces:**
  - For Y_1, and Y_2 in G(p, d), their principle angles are defined as

$$cos(\theta_k) = \max_{u_k \in span(A_1), v_k \in span(A_2)} u_k' v_k$$

$$s.t. \begin{cases} u_k' u_k = 1, v_k' v_k = 1 \\ u_k' u_i = 0, v_k' v_i = 0 \end{cases}$$



  - Where, $\{u_k\}$ and $\{v_k\}$ are called principle dimensions for span(A_1) and span(A_2).

Principle Angles Computing

- **The principle angles between two subspaces:**
  - For A_1, and A_2 in G(p, d), their principle dimensions and angles are computed by SVD:

$$[U, S, V] = SVD(A_1^T A_2)$$

  - Where, $U=[u_1, u_2, \ldots, u_p]$, and $V=[v_1, v_2, \ldots, v_p]$ are the principle angles.
  - The diagonal of S, $[s_1, s_2, \ldots, s_p]$ are the cosine of principle angles,

$$s_k = cos(\theta_k)$$

Subspace Distance on Grassmannian Manifold

- **Subspace distances** **[J. Hamm's Phd thesis]**
  - Projection Distance

    Def:

$$d_{prj}(A_1, A_2) = (\sum_{i=1}^{p} sin^2\theta_i)^{1/2}$$

    Computing:

$$d_{prj}^2(A_1, A_2) = p - \sum_{i=1}^{p} cos^2\theta_i = m - ||A_1' A_2||_F^2$$

  - Binet-Cauchy Distance

    Def:

$$d_{bc}(A_1, A_2) = (1 - \prod_i cos^2\theta_i)^{1/2}$$

    Computing:

$$d_{bc}^2(A_1, A_2) = 1 - \prod_i cos^2\theta_i = 1 - det^2(A_1' A_2)$$

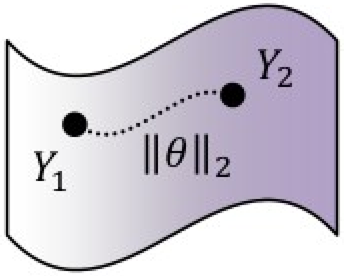
Subspace Distance on Grassmannian Manifold

- **Subspace distances**
  - Arc Distance

    Def:

$$d_{arc}(A_1, A_2) = (\sum_i \theta_i^2)^{1/2}$$

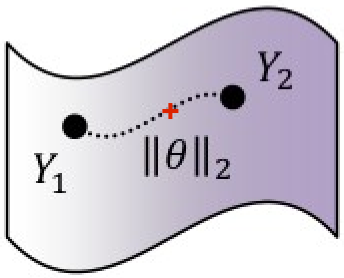
Also known as geodesic distance. It traverse the Grassmannian surface, and two subspace collapse into one, when all principle angles becomes zero.



Weighted Merging of two subspaces

- **What if we need merge two subspaces ?**
  - Motivation:
    - » say if subspace A_1 is best for data set S_1, and subspace A_2 is best for data set S_2, can we find a subspace A_3 that is good for both ?
  - When two subspaces are sufficiently close on Grassmannian manifold, we can approximate this by, $A_3=[t_1, t_2, ...]$

$$t_k = \frac{n_1}{n_1 + n_2} u_k + \frac{n_2}{n_1 + n_2} v_k$$

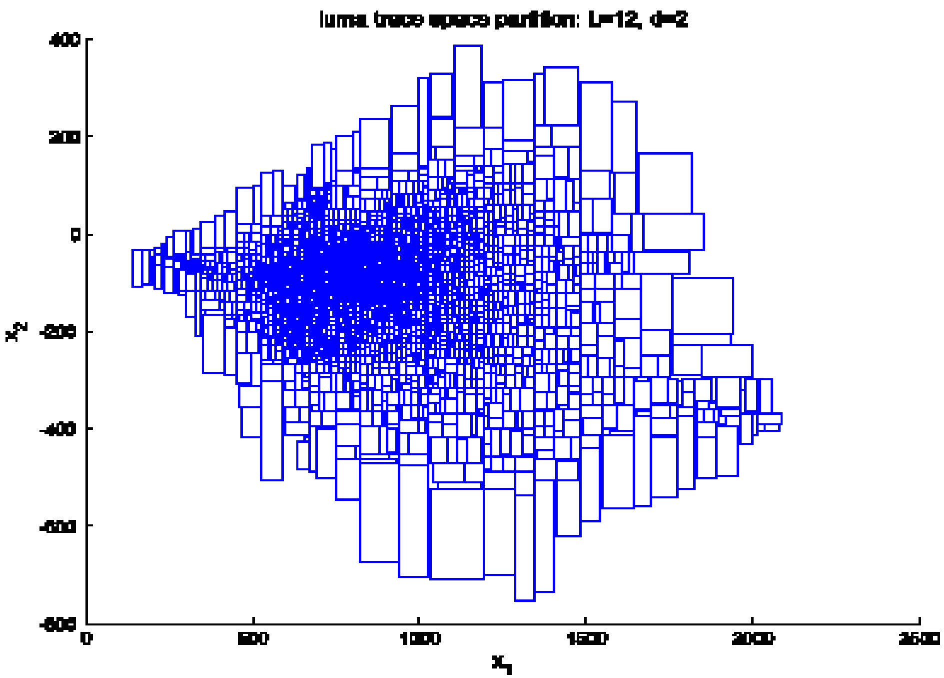


Where $n_{1,2}$ are the size of data set $S_{1,2}$

  - The new sets of basis may not be orthogonal. Can be corrected by Gram-Schmidt orthogonalization.

Judicious Local Models

- **Data Space Partition**
  - Partition the training data set by kd-tree
  - For the kd-tree height of h, we have 2^h local data patch as leaf node
  - For each leaf node data patch $_k$, build a local LDA/LPP/PCA model A_k:



Subspace Index

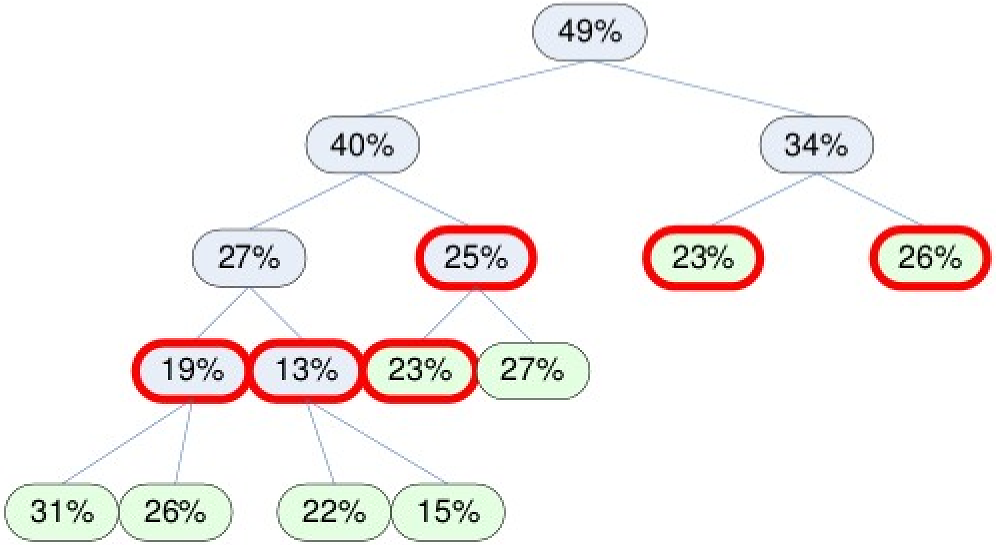
- **Organizing the Subspace Models**
  - For data index of height of h, we have 2^h local models A_k: $k=1..2^h$.
  - For a given probe data point, find its leaf node and associated local model, do identification. Is this good ?
  - No, because
    - » Could be over-fitting, not sure what is the right size local data patch.
    - » Improper neighborhood, probe data points falling on the boundary of leaf node:
  - Build local models at each subtree ?
    - » No, the data partition does not reflect the smooth change of the local models.

Model Hierarchical Tree (MHT)

- **Indexing Subspaces on Grassmannian manifold**
  - It is a VQ like process.
  - Start with a data partition kd-tree, their leaf nodes and associated subspaces $\{A_k\}$, $k=1..2^h$
  - Repeat
    - » Find A_i and A_j, if $d_{arc}(A_i, A_j)$ is the smallest among all, and the associated data patch are adjacent in the data space.
    - » Delete A_i and A_j, replace with merged new subspace, and update associated data patch leaf nodes set.
    - » Compute the empirical identification accuracy for the merged subspace
    - » Add parent pointer to the merged new subspace for A_i and A_j .
    - » Stop if only 1 subspace left.
  - Benefit:
    - » avoid forced merging of subspace models at data patches that are very different, though adjacent.

MHT Based Identification

- **MHT operation**
  - Organize the leaf nodes models into a new hierarchy, with new models and associated accuracy (error rate) estimation
  - When a probe point comes, first identify its leaf nodes from the data partition tree.
  - Then traverse the MHT from leaf nodes up, until it hits the root, which is the global model, and choose the best model along the path for identification

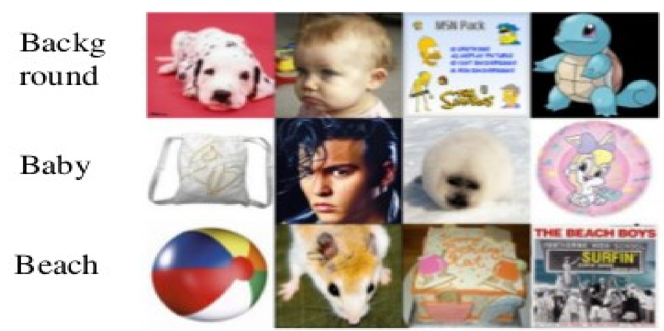


Simulation

- **The data set**
  - MSRA Multimedia data set, 65k images with class and relevance labels:

'Very relevant' samples from three classes: *background, baby* and *beach*

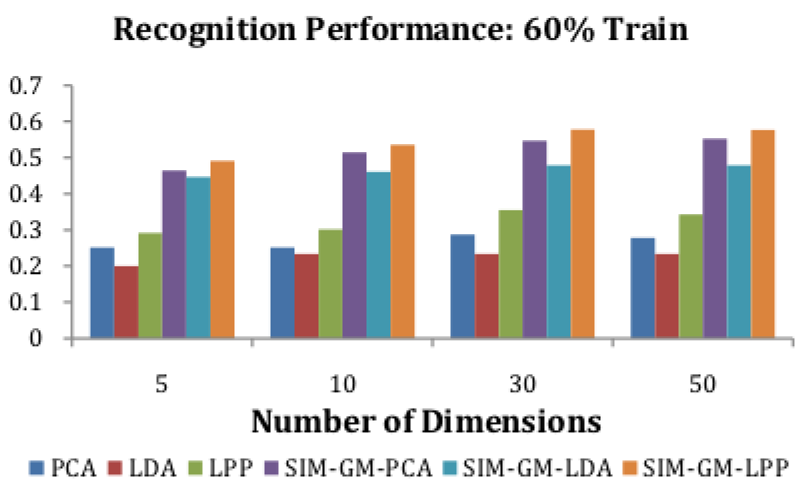
'Relevant' samples from the three classes

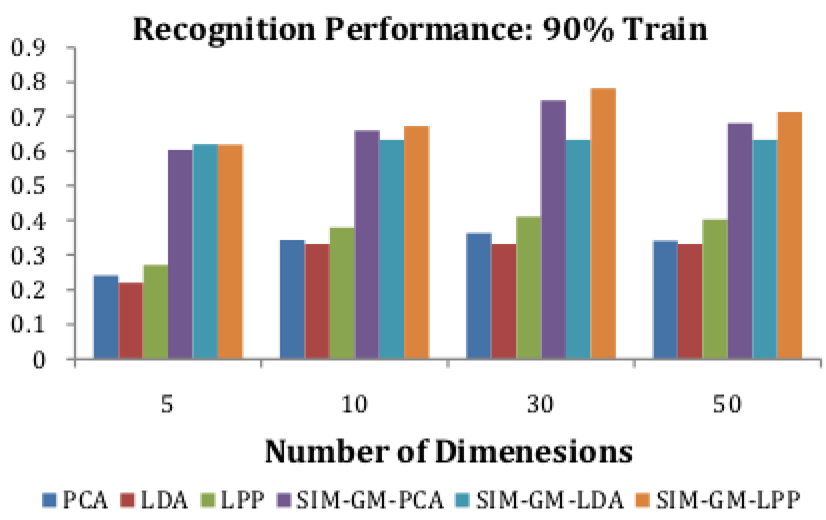


'Irrelevant' samples from the three classes

Simulation

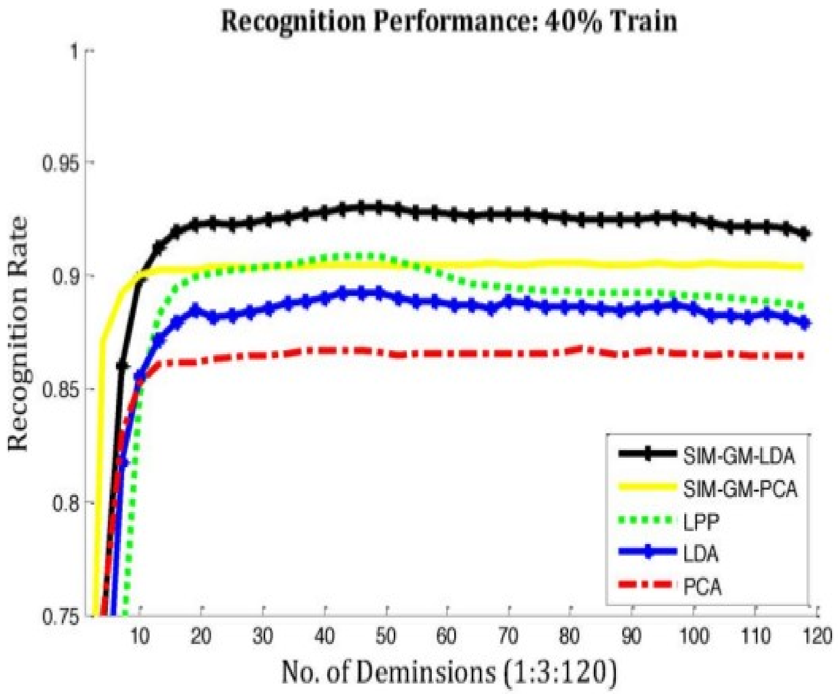
- **Data selection and features**
  - Selected 12 classes with 11k images and use the original combined 889d features from color, shape and texture
  - Performance compared with PCA, LDA and LPP modeling

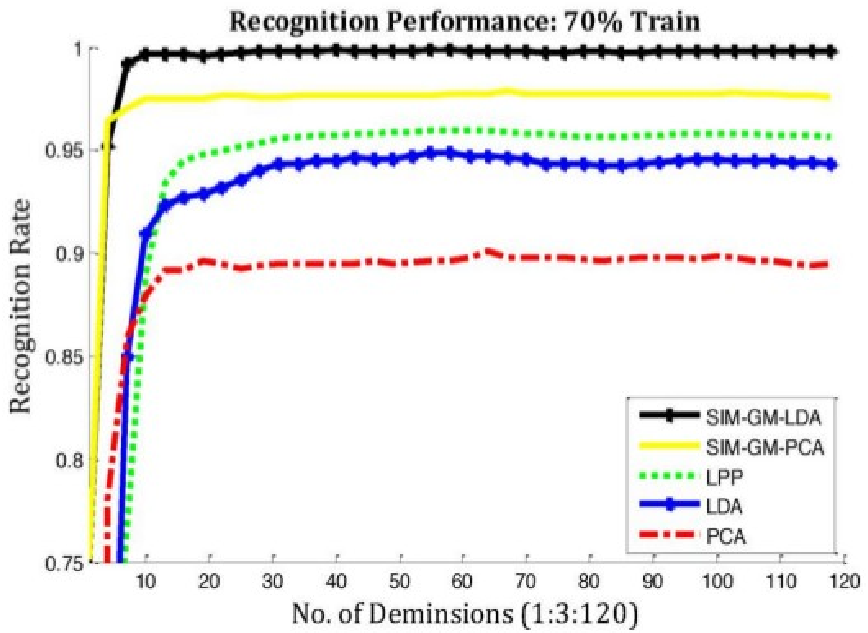




Simulation

- **Face data set**
  - Mixed data set of 242 individuals, and 4840 face images
  - Performance compared with PCA, LDA and LPP modeling





Summary

- **Contributions**
  - The work is a piece-wise linear approximation of non-linear appearance manifold
  - Query driven provide suboptimal performance but still better than a global model.
  - It offers best local models for identification by deriving the subspace structure/index with metrics on Grassmannian manifold
  - Guaranteed performance gains, and the root model degenerates into the global linear model

- **Limitations**
  - Do not have a continuous characterization of Identification error function on the Grassmann manifold.
  - Still heavy on storage cost
  - Need to get more large scale data set to test it.

Summary

- **Future work**
  - Grassmann Hashing – Penalize projection selection with Grassmannian metric, offers performance gains over LSH and spectral hashing.
  - Gradient and Newtonian optimization on Grassmannian manifold.

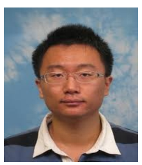
Related papers

- X. Wang, Z. Li, and D. Tao, "Subspace Indexing on Grassmann Manifold for Image Search", IEEE Trans. on Image Processing, vol. 20(9), 2011.
- X. Wang, Z. Li, L. Zhang, and J. Yuan, "Grassmann Hashing for Approx Nearest Neighbour Search in High Dimensional Space", Proc. of IEEE Int'l Conf on Multimedia & Expo (ICME), Barcelona, Spain, 2011.
- H. Xu, J. Wang, Z. Li, G. Zeng, S. Li, "Complementary Hashing for Approximate Nearest Neighbor Search", IEEE Int'l Conference on Computer Vision (ICCV), Barcelona, Spain, 2011.
- Yun Fu, Z. Li, J. Yuan, Ying Wu, and Thomas S. Huang, "Locality vs. Globality: Query-Driven Localized Linear Models for Facial Image Computing," IEEE Transactions on Circuits and Systems for Video Technology (T-CSVT), vol. 18(12), pp. 1741-1752, December, 2008.

Acknowledgement

- **Grants:**
  - The work is partially supported by;
    - » a Hong Kong **RGC** Grant, and
    - » **Microsoft Research Asia** faculty grant.

- **Collaborators:**



» Xinchao Wang, valedictorian of Dept of COMP, HK Polytechnic University, class 2010, now PhD at EPFL

» Dacheng Tao, Professor at Univ of Technology of Sydney.

Q&A

- **Questions please……**

Thanks !